The structure of the protein universe and genome evolution

Eugene V. Koonin, Yuri I. Wolf & Georgy P. Karev

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA (e-mail: koonin@ncbi.nlm.nih.gov)

Despite the practically unlimited number of possible protein sequences, the number of basic shapes in which proteins fold seems not only to be finite, but also to be relatively small, with probably no more than 10,000 folds in existence. Moreover, the distribution of proteins among these folds is highly non-homogeneous — some folds and superfamilies are extremely abundant, but most are rare. Protein folds and families encoded in diverse genomes show similar size distributions with notable mathematical properties, which also extend to the number of connections between domains in multidomain proteins. All these distributions follow asymptotic power laws, such as have been identified in a wide variety of biological and physical systems, and which are typically associated with scale-free networks. These findings suggest that genome evolution is driven by extremely general mechanisms based on the preferential attachment principle.

The distribution of matter and energy in the Universe provides cosmologists with the principal source of information on the evolution of our planet, including its earliest stages. In particular, the discovery of the uniformly distributed background microwave radiation is the main proof of the Big Bang model of the Universe's origin (for example, see refs 1, 2). In a somewhat loose but perhaps appropriate analogy, structural biologists often speak of the 'protein universe', meaning the totality of all possible proteins[1–3]. The total number of possible protein sequences (that is, the size of the protein universe) is, for all practical purposes, infinite. Assuming an average protein length of 200 amino acids, there can be 20^{200} different protein sequences, a number that is much greater than, for example, the number of electrons in our (physical) Universe.

Our current theoretical understanding of protein folding is insufficient to estimate the total possible number of protein structures, but it too is likely to be vast. Obviously, only a minuscule fraction of the potential sequence space is populated by real protein sequences, but the number of unique sequences encoded in actual genomes is likely to be substantial. For example, assuming there are 10 million species on Earth and the genome of each species consists of 5,000 genes (an intermediate number between prokaryotes and eukaryotes), there are 5×10^{10} unique protein sequences. Although this quantity is negligible compared to the vast sequence space, it still is several orders of magnitude greater than that contained in today's databases. A question of fundamental and practical interest is how these sequences are distributed in the sequence and structure spaces.

The protein universe is an abstraction, however useful. In reality, all proteins are, of course, encoded in genes, which belong to particular genomes. Quantitative and qualitative analysis of the projections of the protein universe on genomes from a diverse range of organisms might reveal important aspects of the evolution of both genomes and proteins.

Distribution of protein families and protein folds

That the population of the protein universe is not distributed randomly is obvious from the existence of homologous genes and proteins. However, to extract any useful information from this distribution, it needs to be explored in quantitative detail, which can be done only within the framework of a hierarchical taxonomy of proteins. Margaret Dayhoff's group introduced the notions of protein family and superfamily in the 1960s as part of their effort to understand protein evolution and simultaneously create a well-organized protein database[4,5] (later known as the Protein Identification Resource). A family was defined as a group of (closely) related sequences, and superfamilies encompassed two or more related families.

By the mid-1990s, a more elaborate and coherent taxonomy of protein domains had been developed, largely through the efforts of Murzin and colleagues, who constructed the SCOP classification of protein structures, and Thornton and colleagues, who produced the CATH database dedicated to the same goal[1,6–9]. The top levels of the hierarchy are defined by the three-dimensional structure, whereas lower taxa are identified on the basis of sequence similarity and functional considerations (Table 1). Exact criteria of topological similarity, which is necessary and sufficient to assign two protein structures to the same fold, or the level of sequence similarity that defines a superfamily or a family, have yet to be determined in full. Nevertheless, there is a wide agreement both on the general principles of classification and on the taxonomic assignments of most proteins[6,8,10,11].

At this point, it is important to introduce the fundamental notion of protein domain, which is the foundation of at least the top levels of the protein taxonomy. In structural biology, a domain is defined as a distinct, compact and stable protein structural unit that folds independently of other such units[12]. Often, however, domains are characterized differently — as distinct regions of protein sequence that are highly conserved in evolution. As the hierarchy of protein classification evolved into a combination of structural- and sequence-based approaches, the notions of structural and 'homology' domains also tended to blend into one concept. The salient features of structural domains (that is, independent folding and stability) conduce them to become distinct evolutionary units, which exist as stand-alone proteins or as parts of various domain architectures in multidomain proteins. There are exceptions to this generalization, one

Table 1 **Hierarchical classification of proteins**

Category	Example	Definition, criteria or main features
Structural class	α/β	Overall composition of structural elements. No evolutionary relationship.
Fold	TIM barrel	Topology of the folded protein backbone. Monophyletic origin?
Superfamily	Aldolase	Recognizable sequence similarity (at least a conserved motif); conservation of basic biochemical properties. Monophyletic origin.
Family	Class I aldolase	Significant sequence similarity; conservation of biochemical activity.
Group of orthologues (COG)	2-keto-3-deoxy-6-phospho-gluconate aldolase	Orthologous relationships within the given set of species; conservation of the biochemical activity and, most often, also the biological function.
Lineage-specific expansion	PA3131 and PA3181 in *Pseudomonas aeruginosa*	Paralogues originating from a lineage-specific duplication; possible functional specialization.

being where two structural domains comprise a seemingly inseparable evolutionary unit (a 'homology domain'). But whenever this situation is observed, stand-alone versions or new multidomain architectures of the respective domains are usually discovered eventually; this is supported by numerous observations made in the context of recent genome analyses (for example, see refs 13, 14).

There is no doubt that protein families and superfamilies are monophyletic, that is, they derive from a common ancestor. In contrast, monophyly of protein folds, as opposed to folds originating by convergence from unrelated ancestors, remains an issue of debate. It seems that, for most folds (with the possible exception of some of the most diverse 'superfolds'), similarity goes beyond the topology of the protein backbone. Often, the basic physicochemical interactions and the associated structural and sequence motifs are conserved throughout a fold (for example, the P-loop in the eponymous ATP/GTPase fold[15,16]), or even across fold boundaries (for example, the phosphate-binding loop in Rossmann-type nucleotide-binding domains[17]). Perhaps more important, on numerous occasions, the same activity and/or function is performed by two or more unrelated folds in different organisms or in different cellular systems in the same organism[18,19]. Taken together, these observations seem to argue against convergence as the prevalent force in the evolution of protein folds and suggest that most, if not all, protein folds are monophyletic. However, the possibility of multiple, convergent origins still might be considered for some common folds with a relatively simple, symmetric topology, such as TIM barrels (named after the structure of the glycolytic enzyme triosephosphate isomerase) or β-propellers.

Protein families consist of related 'individuals', each of which is a set of orthologues, or proteins related by vertical descent (according to the classification of homologues proposed by Walter Fitch[20,21]). Clusters of orthologous groups of proteins (COGs) typically occupy a unique functional niche, which remains the same in different, even phylogenetically distant organisms, except for lineage-specific expansions of proteins within a COG[22,23]. These expansions that result from relatively recent duplications are prominent in genomes, particularly in eukaryotes[24–26]. On many occasions, there is a plausible connection between the lineage-specific proliferation of a particular family and specific adaptations characteristic of the given group of organisms. The relationships between distinct COGs within a family (as well as between families within a superfamily and, most likely, between superfamilies within a fold) represent paralogy, that is, origin from an ancestral duplication[20,21,27]. Paralogous COGs within a family tend to have different biological functions, although, in many cases, they have identical or similar biochemical activities.

Early sequence and structure databases were severely biased, primarily because of overrepresentation of sequences of well-characterized proteins and gross under-representation of uncharacterized ones. Growth of the databases, especially with the advent of high-throughput genome sequencing, eliminated much of this sampling bias. By mid-1990, it became clear that the distribution of protein domains among folds, superfamilies and families was extremely uneven — most taxa consisted of a small number of members and only a few were highly abundant. Rigorous application of sampling theory ruled out sampling bias as the principal contribution to the observed distribution[28,29].

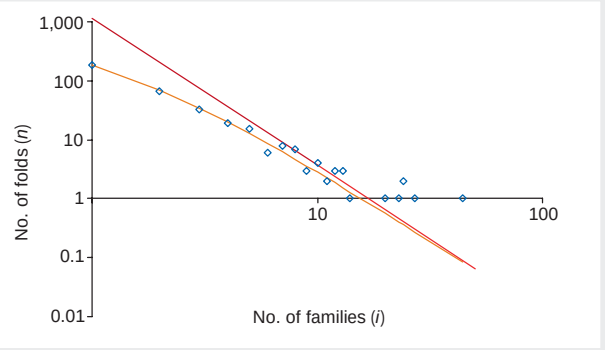

Figure 1 Double-logarithmic plot of the distribution of protein folds by the number of families. The sequences from the Structural Classification of Proteins (SCOP) 1.39 database were analysed as described in ref. 71. The best fit is defined by the equation $n = 1{,}165(i + 1.1)^{-2.5}$. The red line shows the power-law asymptotic.

As the sampled fraction of the protein universe increased, more reliable estimates of the overall variety of proteins became feasible. In contrast to the earlier assessments, which relied largely on the rate of discovery of new protein families[4,30,31], these studies used the observed distributions of families among folds to extrapolate the total numbers, taking the sampling process into account. Depending on the assumptions and methods used, the estimates of the total number of existing protein folds produced by different researchers varied substantially, from ~650 to ~10,000 (refs 29, 32–37). But examination of the distribution of folds by the number of protein families (Fig. 1) indicates that, in one sense, the discrepancy between these estimates might be of little consequence. This distribution contains a small number of folds with a large number of families (mostly well-known superfolds, such as P-loop NTPases, the Rossmann fold or TIM barrels) and an increasing number of folds that consist of a small number of families. By far the largest size class consists of the 'unifolds'[37], each including one family, often just one COG. Thus, it seems certain that the great majority of protein families belong to ~1,000 common folds. What is still in dispute is the number of unifolds that encompass the rest of the proteins. Approximately one half of the common folds are currently represented by at least one experimentally determined structure, which means that coarse-grain mapping of the protein universe is already at an advanced stage.

Power laws and models of genome evolution

Mathematically, the distribution of protein folds by the number of constituent families has been approximated by a power law, that is, $f(i) \sim i^{-\gamma}$ where $f(i)$ is the frequency of folds that include exactly i families and γ is a parameter that typically assumes values between 1 and 3 (ref. 34). More precisely, the fold–family distribution fits the so-called generalized Pareto function $f(i) \sim (i + a)^{-\gamma}$, where a is an additional parameter, with the power law fitting asymptotically with the increase of i (Fig. 1). Remarkably, the same function, up to the parameters, fits the distribution of protein domain families by the number of members in each analysed genome, as recently shown by Kuznetsov[38] and by ourselves[39,40] (Fig. 2). These distributions, along with the distributions of other genome-associated quantities (for example, the number of pseudogenes per gene family), have been previously approximated with power laws, first in the pioneering work of Huynen and Van Nimwegen[41] and subsequently in detailed studies by Gerstein and colleagues[42–44].

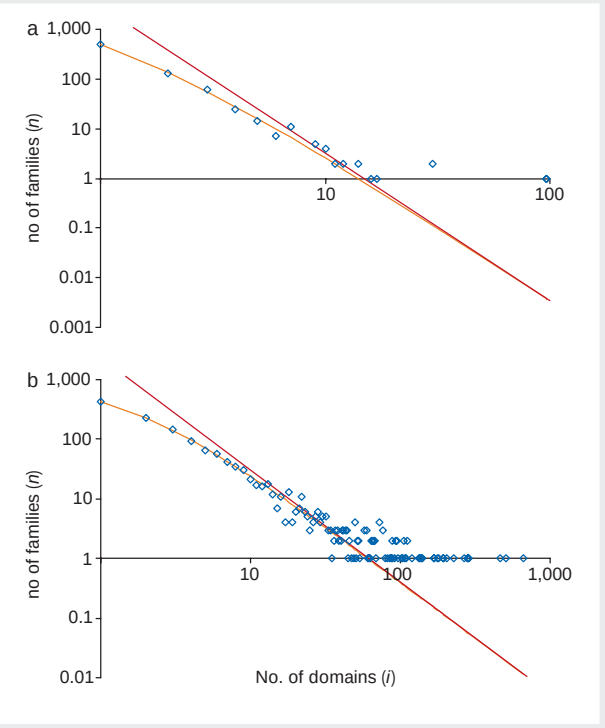

Figure 2 Double-logarithmic plots of the size distribution of protein domain families in genomes. **a**, The hyperthermophilic bacterium *Thermotoga maritima*; $n = 2{,}972(i + 0.8)^{-3.0}$. **b**, The nematode worm *Caenorhabditis elegans*; $n = 2{,}395(i + 1.5)^{-1.9}$. The domains were identified using the Conserved Domain Database (CDD) collection of position-specific scoring matrices and the reversed position-specific (RPS)-BLAST program[77]. Data are from ref. 39; the red line shows the power-law asymptotic

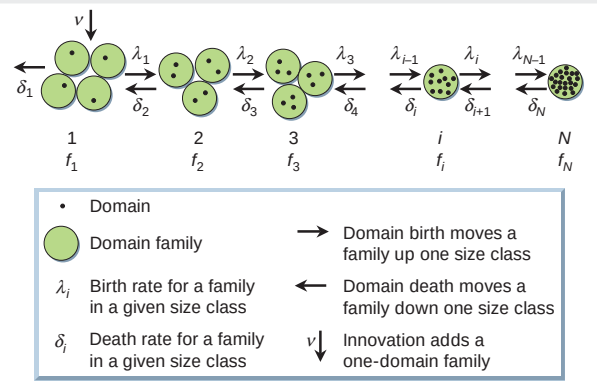

Figure 3 A general scheme of a birth, death and innovation model of genome evolution.

As demonstrated by Barabasi and colleagues and by several other researchers, power laws describe the distribution of various quantities in numerous biological, physical and social contexts; such distributions can seem to be fundamentally different (for example, the number of links between documents in the Internet, the population of towns and the number of reactions in which a given metabolite is involved)[45–50]. Zipf's law, which describes the frequency distribution of words in texts[51], and the Pareto principle, describing the distribution of people by wealth[52], are in this category. These distributions have specific mathematical properties related to those of so-called scale-free networks, that is, networks in which the frequency distribution of node degrees (the number of nodes to which a given node is connected) follows a power law[47,48]. In particular, the network of metabolic reactions in any organism is a scale-free network with a distinct hierarchical structure[50,53] and protein–protein interaction networks have similar properties[54].

The wide spread of power distributions and scale-free networks in nature and society suggests that similar laws might govern evolution in a variety of diverse systems. The general pattern of network evolution that ensures scale-free behaviour is preferential attachment, where the probability of a node acquiring a new connection is proportional to the degree (the number of connections) of that node. Metaphorically, this can be described as a situation in which 'the rich get richer' or, from a selectionist perspective, 'the fit get fitter'[47].

Returning to protein domains, there seems to be at least three (not necessarily exclusive) ways to explain the emergence of power laws and related highly skewed distributions of the fold and family sizes in the protein universe and in individual genomes. The 'designability' hypothesis, favoured by some structural biologists, postulates that certain folds serve as attractors in the space of protein structures because of their topological properties (for example, the highly abundant TIM-barrel fold is a uniquely symmetrical construction). As a result, many unrelated sequences tend to adopt the same few folds. Interestingly, the simulated designability distributions analysed by Wingreen and colleagues[55,56] appear to be similar to the empirical distributions of domain family sizes described by Gerstein and co-workers[42,44], Kuznetsov[38] and ourselves[39]. However, given the above argument against a convergent origin of most folds, designability does not seem to be a likely general explanation for the observed preferential attachment or, more precisely, preferential proliferation of domains in protein evolution.

A straightforward selectionist interpretation holds that certain biochemical activities (for example, nucleoside 5′-triphosphate hydrolysis), being particularly common and important in cellular biochemistry, are in greater demand than other, highly specialized ones, which leads to preferential proliferation of the respective protein families. Again, the weakness of this argument is that the same activity is often embodied in two or more distinct domains, which tend to differ substantially in abundance[18].

Finally, domain birth and death models developed by Gerstein and co-workers[42], Rzhetsky and Gomez[57] and ourselves[39], which originate from the classic analysis of Yule[58], completely disregard the protein identity, but give rise to equilibrium distributions of domain family sizes that show an excellent fit to the observed ones. These models typically include the elementary processes of family growth via domain birth (duplication), domain death as a result of inactivation and loss, and innovation or emergence of a new family (for example, through extensive modification of a member of an existing family, horizontal gene transfer or even origin of a new protein from non-coding sequence) (Fig. 3).

We recently explored the behaviour of these birth, death and innovation models (BDIMs) in detail, both analytically and by computer simulation; this analysis seems to lead to non-trivial conclusions on genome evolution[39]. First, it was shown that, using BDIMs, an equilibrium distribution of domain family sizes is reached exponentially fast during evolution from any initial conditions. Specifically, $|f_i(t) - f_i| \sim e^{-kt}$, where $f_i(t)$ is the frequency of a given family at time t and f_i is the equilibrium frequency. Thus, any perturbation in genome evolution, which involves changes in the parameters of birth, death or innovation, rapidly relaxes to a new stationary state. Accordingly, the mode of evolution depicted by BDIMs is most compatible with the punctuated equilibrium notion of genome evolution[59]. By this model, long periods of stasis are punctuated by relatively brief bursts of evolutionary activity, which involve rapid proliferation and elimination of gene families as well as 'invention' and acquisition of new ones.

Second, BDIMs result in different shapes of equilibrium distributions of family sizes depending on how precisely the birth rate is balanced by the death rate. The power law appears as an asymptotic in a certain, specific subclass of BDIM, in which the death rate

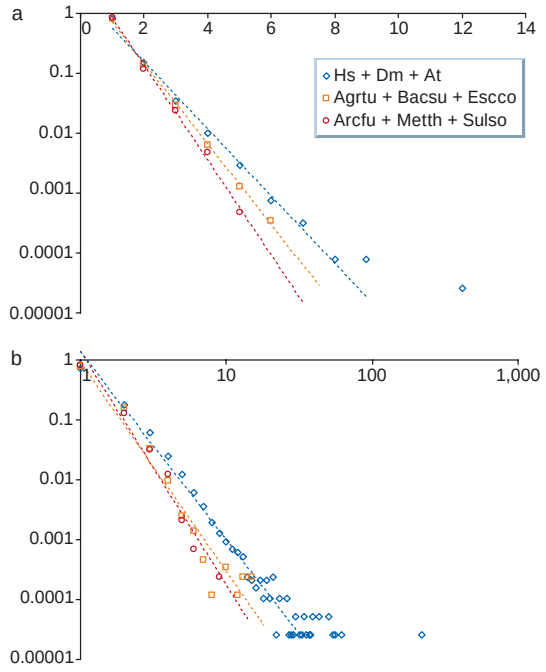

Figure 4 Distributions of the number of domains in proteins from the three primary kingdoms of life. **a**, Repeats of the same domain in a single polypeptide excluded. The plot is in semi-logarithmic scale. **b**, Repeats of the same domain in a single polypeptide included. The plot is in double logarithmic scale. The data and methods used for generating this plot were the same as in Fig. 2. Eukaryotes: Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; At, *Arabidopsis thaliana*. Bacteria: Agrtu, *Agrobacterium tumefaciens*; Bacsu, *Bacillus subtilis*; Escco, *Escherichia coli*. Archaea: Arcfu, *Archaeoglobus fulgidus*; Metth, *Methanothermobacter thermoautotrophicus*; Sulso, *Sulpholobus solfataricus*.

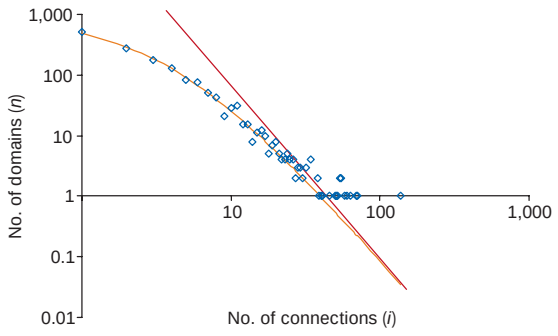

Figure 5 Double-logarithmic plot of the distribution of protein domains by the number of links in multidomain proteins. The number of links is the number of different domains with which the given domain combines in multidomain proteins. The combined data from seven analysed bacterial genomes, three archaeal genomes and six eukaryotic genomes were the same as in Fig. 2, except that several domains that showed artificially high numbers of connections because of their biased amino acid composition were removed manually. The best fit is given by the equation $n = 46{,}815(i + 4.0)^{-2.9}$.

approaches the birth rate for large families, but is considerably greater than the birth rate for small families. These models accurately describe the distributions of domain family size for all analysed genomes, whereas straightforward approximation with a power law does not fit the data nearly as well (Fig. 2).

Finally, analysis of BDIMs shows that the innovation rate, which is required to offset the stochastic loss of low-copy families, has to be relatively high and, at least in small, prokaryotic genomes, comparable to the overall intra-genomic duplication (birth) rate. This supports, from a somewhat unexpected angle, the key role of horizontal gene transfer in prokaryotic evolution that has been suggested by numerous observations made during genome comparisons[60–64].

The evolutionary models described here ignore completely the individuality of gene families and the selective forces that make some of them expendable and others indispensable. Despite this obvious over-simplification, BDIMs accurately reproduce the observed family size distributions, suggesting that genome evolution might be largely a stochastic process, which is only modulated by selection.

Paradoxes of multidomain networks

Protein domains often combine to form multidomain architectures. Analysis of such architectures can be extremely helpful for predicting functions of uncharacterized domains and proteins in a 'guilt by association' approach (also called the 'Rosetta Stone' principle), which is based on the assumption that physical fusion of two domains implies a functional link[65–68]. Indeed, multidomain proteins have critical roles in all living cells, as they provide effective links between different functional systems. Because of this ability, complex multidomain architectures are particularly characteristic of various signalling systems.

There seems to be a connection between the propensity of protein domains to form multidomain architectures and the organismic complexity. Specifically, in many orthologous sets of eukaryotic proteins, such as chromatin-associated transcription factors, a distinct trend, which we dubbed 'domain accretion', can be traced towards increased complexity of domain architectures in more complex organisms[69]. Because proteins form complex networks, even a modest increase in the number of domains in interacting partners could translate into numerous new interactions, which probably contributes to the solution of the apparent paradox of 'too few' genes in complex organisms[70].

Given the involvement of multidomain proteins in a variety of cellular functions, we might expect that natural selection should favour their formation to the extent that multidomain architectures would be over-represented with respect to single-domain proteins, especially in complex eukaryotes. However, quantitative analysis does not seem to support this conclusion. Instead, the distribution of proteins by the number of different domains (with multiple occurrences of the same domain in a given protein excluded from the analysis) shows an excellent fit to an exponent[71] (Fig. 4a). This type of distribution is compatible with a random recombination (joining and breaking) model of evolution of multidomain architectures.

Notably, however, the slopes of the curves in Fig. 4a differ significantly for archaea, bacteria and eukaryotes, indicating that the fraction of multidomain proteins or, in terms of the random model, the likelihood of domain joining increases in the order: archaea < bacteria <eukaryotes. The under-representation of multidomain proteins in archaea compared to the other two primary kingdoms of life might be related to the low stability of large proteins in the hyperthermophilic habitats of most archaeal species. The excess of multidomain proteins in eukaryotes is not unexpected given the observations on domain accretion; furthermore, the right tail of the eukaryotic distribution shows a deviation from the exponent caused by the presence of several proteins with a large number of domains (Fig. 4a). When repeats of the same domain in a single polypeptide chain are added to the mix, the distribution changes and is best approximated by a generalized Pareto function (Fig. 4b). In light of the above, this finding does not seem unexpected: evolution of repeats is likely to follow a BDIM scenario, with tandem duplication and elimination as the main underlying processes, rather than a random joining–breaking model, which seems to apply to combinations of different domains.

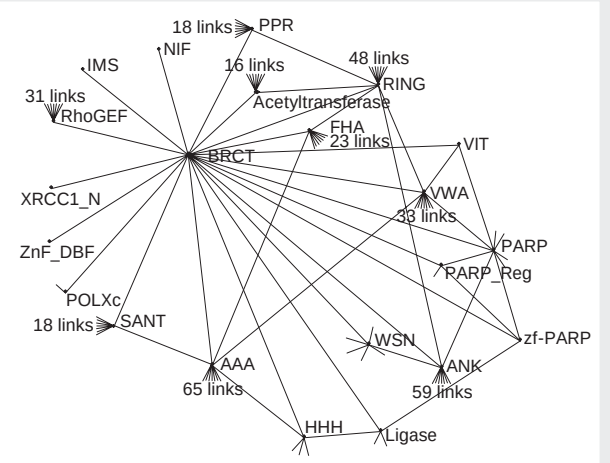

Figure 6 A fragment of the network of multidomain connections. All the connections of the BRCT (BRCA1 C-terminal) domain and those between its partners are shown; the number of outgoing connections is also indicated for all domains other than BRCT.

Table 2 **Top multidomain connections in eukaryotes, bacteria and archaea**

Eukaryotes		Bacteria		Archaea	
NC	Domain	NC	Domain	NC	Domain
133	Ser/Thr protein kinase	28	AAA ATPase	18	AAA ATPase
69	PH	23	Receiver domain	9	His-kinase-type ATPase
63	Ankyrin repeat	20	His-kinase-type ATPase	9	CBS
60	RRM	18	GAF	9	D-Ala-D-Ala ligase
58	Immunoglobulin	15	CBS	9	4Fe-4S ferredoxin
55	PHD finger	15	PAS	8	DEXD helicase
55	PDZ	14	His-kinase phosphoacceptor	8	PAS
54	SH3	14	HAMP	8	PAC
54	RING finger	14	FMO-like	8	FMO-like
52	C2	14	ACT	7	Receiver

Abbreviations and acronyms: NC, Number of connections; PH, plekstrin homology domain; RRM, RNA-recognition motif; PHD, plant homeodomain (-associated finger); PDZ, postsynaptic density protein-95/discs large/zo-1 domain; SH3 Src homology 3 domain; FMO, flavin-dependent monooxygenase; GAF, CBS, PAS, PAC, HAMP, ACT are small-molecule-binding domains involved in various forms of signal transduction and allosteric regulation of enzymatic activity.

The above analysis does not tell us anything about the propensity of individual domains to form multidomain architectures, and these propensities differ widely. In an already familiar pattern, the distribution of the number of multidomain architectures in which a domain is involved roughly follows a power law, as demonstrated by Wuchty[72] and by Teichmann and co-workers[73]. More precisely, this distribution is described by a generalized Pareto function that we have already encountered in other contexts (Fig. 5). Thus, a small number of domains are hubs of multidomain connections that hold together cellular interaction networks. Although evolution of multidomain proteins containing different domains seems to occur primarily via random processes of joining and breaking (Fig. 4a), the fit (to form functionally advantageous multidomain architectures) still get fitter.

The network of multidomain connections for a moderately linked hub, the BRCT (BRCA1 C-terminal) domain, which is an important adaptor in eukaryotic cell-cycle checkpoints and DNA repair[74,75], is shown in Fig. 6. Notably, some of the domains linked to BRCT, such as RING (involved in ubiquitin-dependent cascades) and FHA (implicated in various signal-transduction pathways) are important hubs themselves. Table 2 shows the top multidomain connectors for bacteria, archaea and eukaryotes. Remarkably, the lists for the two prokaryotic kingdoms have five domains in common, whereas the eukaryotic list is completely different. Not unexpectedly, however, all three sets are dominated by domains that are involved in various forms of signal transduction and regulation of enzymatic activity.

Perspectives

The protein universe is extremely unevenly populated, with most proteins concentrated in a relatively small number of major clusters, the common folds and superfolds. This highly skewed distribution of proteins among folds should enable structural genomics research programmes to complete a preliminary tour of the most important part of the protein universe within the next few years[76], although many rare folds are likely to remain uncharacterized for much longer.

Projection of the structure of the protein universe on genomes and quantitative analysis of the outcome seems to result in some unexpected insights into general principles of genome evolution. Remarkably, the size distributions of folds for the explored part of the protein universe and of domain families for all analysed genomes, as well as the distribution of the number of domain connections in multidomain architectures, are all described by the same type of mathematical functions, in which the power law appears as an asymptotic. This suggests that extremely general mechanisms of evolution, apparently based on the preferential attachment (proliferation) principle, are at work in all these contexts.

With respect to domain families, these principles have already been detailed in plausible, even if oversimplified, models of genome evolution based on the elementary processes of birth, death and innovation. Similar models could potentially be developed for other situations, such as the connections between domains in multidomain networks, as well as networks of protein–protein interactions and metabolic reactions. Evolutionary modelling certainly needs to be made more realistic by including additional parameters, particularly those associated with purifying and positive selection. It seems reasonable to hope that further quantitative analysis of the structure of the protein universe and its projections on diverse genomes ushers a qualitatively new understanding of the evolution of life in a not so remote future. □

doi:10.1038/nature01256

1. Holm, L. & Sander, C. Mapping the protein universe. *Science* **273,** 595–603 (1996).
2. Zhang, C. & DeLisi, C. Protein folds: molecular systematics in three dimensions. *Cell. Mol. Life Sci.* **58,** 72–79 (2001).
3. Rost, B. Did evolution leap to create the protein universe? *Curr. Opin. Struct. Biol.* **12,** 409–416 (2002).
4. Dayhoff, M. O. The origin and evolution of protein superfamilies. *Fed. Proc.* **35,** 2132–2138 (1976).
5. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. Establishing homologies in protein sequences. *Methods Enzymol.* **91,** 524–545 (1983).
6. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).
7. Murzin, A. G. Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6,** 386–394 (1996).
8. Orengo, C. A. *et al.* CATH—a hierarchic classification of protein domain structures. *Structure* **5,** 1093–1108 (1997).
9. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307,** 1113–1143 (2001).
10. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30,** 264–267 (2002).
11. Orengo, C. A. *et al.* The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* **2,** 11–21 (2002).
12. Branden, C.-I. & Tooze, J. *Introduction to Protein Structure* (Garland Publishing, New York, 1999).
13. Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* **30,** 1427–1464 (2002).
14. Anantharaman, V., Koonin, E. V. & Aravind, L. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.* **307,** 1271–1292 (2001).
15. Saraste, M., Sibbald, P. R. & Wittinghofer, A. The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15,** 430–434 (1990).
16. Koonin, E. V. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J. Mol. Biol.* **229,** 1165–1174 (1993).
17. Aravind, L., Mazumder, R., Vasudevan, S. & Koonin, E. V. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **12,** 392–399 (2002).
18. Galperin, M. Y., Walker, D. R. & Koonin, E. V. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8,** 779–790 (1998).
19. Martin, A. C. *et al.* Protein folds and functions. *Structure* **6,** 875–884 (1998).
20. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19,** 99–113 (1970).

21. Fitch, W. M. Homology a personal view on some of the problems. *Trends Genet.* **16,** 227–231 (2000).
22. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
23. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
24. Jordan, I. K., Makarova, K. S., Spouge, J. L., Wolf, Y. I. & Koonin, E. V. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* **11,** 555–565 (2001).
25. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314,** 1041–1052 (2001).
26. Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12,** 1048–1059 (2002).
27. Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278,** 609–614 (1997).
28. Alexandrov, N. N. & Go, N. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci.* **3,** 866–875 (1994).
29. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372,** 631–634 (1994).
30. Zuckerkandl, E. The appearance of new structures and functions in proteins during evolution. *J. Mol. Evol.* **7,** 1–57 (1975).
31. Chothia, C. One thousand families for the molecular biologist. *Nature* **357,** 543–544 (1992).
32. Zhang, C. T. Relations of the numbers of protein sequences, families and folds. *Protein Eng.* **10,** 757–761 (1997).
33. Wang, Z. X. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng.* **11,** 621–626 (1998).
34. Zhang, C. & DeLisi, C. Estimating the number of protein folds. *J. Mol. Biol.* **284,** 1301–1305 (1998).
35. Govindarajan, S., Recabarren, R. & Goldstein, R. A. Estimating the total number of protein folds. *Proteins* **35,** 408–414 (1999).
36. Wolf, Y. I., Grishin, N. V. & Koonin, E. V. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299,** 897–905 (2000).
37. Coulson, A. F. & Moult, J. A unifold, mesofold, and superfold model of protein fold use. *Proteins* **46,** 61–71 (2002).
38. Kuznetsov, V. A. in *Computational and Statistical Approaches to Genomics* (eds Zhang, W. & Shmulevich, I.) 125–171 (Kluwer, Boston, 2002).
39. Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezovskaya, F. S. & Koonin, E. V. in *Computational Genomics: from Sequence to Function* (eds Galperin, M. Y. & Koonin, E. V.) (Horizon, Amsterdam, in the press).
40. Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezovskaya, F. S. & Koonin, E. V. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol. Biol.* (in the press).
41. Huynen, M. A. & van Nimwegen, E. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15,** 583–589 (1998).
42. Qian, J., Luscombe, N. M. & Gerstein, M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313,** 673–681 (2001).
43. Harrison, P. M. & Gerstein, M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318,** 1155–1174 (2002).
44. Luscombe, N., Qian, J., Zhang, Z., Johnson, T. & Gerstein, M. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol.* **3,** research0040.1–0040.7 (2002).
45. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
46. Bilke, S. & Peterson, C. Topological properties of citation and metabolic networks. *Phys. Rev. E* **64,** 036106-1–036106-5 (2001).
47. Barabasi, A. L. *Linked: The New Science of Networks* (Perseus, New York, 2002).
48. Albert, R. & Barabasi, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74,** 47–97 (2002).
49. Gisiger, T. Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biol. Rev. Camb. Phil. Soc.* **76,** 161–209 (2001).
50. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).
51. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Boston, 1949).
52. Pareto, V. *Cours d'Economie Politique* (Rouge et Cie, Paris, 1897).
53. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297,** 1551–1555 (2002).
54. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
55. Li, H., Helling, R., Tang, C. & Wingreen, N. Emergence of preferred structures in a simple model of protein folding. *Science* **273,** 666–669 (1996).
56. Li, H., Tang, C. & Wingreen, N. S. Are protein folds atypical? *Proc. Natl Acad. Sci. USA* **95,** 4987–4990 (1998).
57. Rzhetsky, A. & Gomez, S. M. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* **17,** 988–996 (2001).
58. Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. *Phil. Trans. R. Soc. Lond. B* **213,** 21–87 (1924).
59. Gould, S. J. *The Structure of Evolutionary Theory* (Harvard Univ. Press, Cambridge, MA, 2002).
60. Doolittle, W. F. Lateral genomics. *Trends Cell Biol.* **9,** M5–M8 (1999).
61. Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284,** 2124–2129 (1999).
62. Doolittle, W. F. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* **14,** 307–311 (1998).
63. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55,** 709–742 (2001).
64. Ragan, M. A. Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.* **11,** 620–626 (2001).
65. Marcotte, E. M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285,** 751–753 (1999).
66. Enright, A. J., Illopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402,** 86–90 (1999).
67. Galperin, M. Y. & Koonin, E. V. Who's your neighbor? New computational approaches for functional genomics. *Nature Biotechnol.* **18,** 609–613 (2000).
68. Aravind, L. Guilt by association: contextual information in genome analysis. *Genome Res.* **10,** 1074–1077 (2000).
69. Koonin, E. V., Aravind, L. & Kondrashov, A. S. The impact of comparative genomics on our understanding of evolution. *Cell* **101,** 573–576 (2000).
70. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
71. Wolf, Y. I., Brenner, S. E., Bash, P. A. & Koonin, E. V. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9,** 17–26 (1999).
72. Wuchty, S. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* **18,** 1694–1702 (2001).
73. Apic, G., Gough, J. & Teichmann, S. A. An insight into domain combinations. *Bioinformatics* **17**(Suppl. 1), S83–S89 (2001).
74. Bork, P. *et al.* A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J.* **11,** 68–76 (1997).
75. Derbyshire, D. J. *et al.* Crystal structure of human 53BP1 BRCT domains bound to p53 tumour suppressor. *EMBO J.* **21,** 3863–3872 (2002).
76. Vitkup, D., Melamud, E., Moult, J. & Sander, C. Completeness in structural genomics. *Nature Struct. Biol.* **8,** 559–566 (2001).
77. Marchler-Bauer, A. *et al.* CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30,** 281–283 (2002).

Acknowledgements

We thank A. Panchenko and S. He (NCBI) for help with the use of the Conserved Domain Database, and A. Rzhetsky and V. Kuznetsov for helpful discussions.